Towards Measuring the Security Quality of Fido2 Deployments at Scale

Macha Krumm
University of Liège
Liège, Belgium
florian.dekinder@uliege.be

Florian Dekinder
University of Liège
Liège, Belgium
florian.dekinder@uliege.be

Benoit Donnet
University of Liege
Liège, Belgium
benoit.donnet@uliege.be

CCS Concepts

• **Security and privacy** → *Web application security*; *Browser security*.

Keywords

Atlas, Fido2, WebAuthn, signals, authentication classification, Fido2 challenge

1 Problem

Fido2/WebAuthn [3, 6] is rapidly being deployed across the Web as the phishing-resistant successor to passwords, yet — unlike most Internet protocols whose correctness can be inferred from passive traffic analysis or protocol conformance testing — its real-world security properties are only observable by actively probing the authentication flow of individual websites at scale. The security of a Fido2 deployment hinges on network-level artifacts exchanged between client and server: the entropy and freshness of the cryptographic challenge, its potential reuse across independent sessions, and the policy fields negotiated during the handshake. Measuring these properties across the Web requires infrastructure and techniques squarely rooted in Internet measurement: automated, non-intrusive probing at scale, careful handling of anti-bot defenses that increasingly shape what is observable from outside a network, and statistically sound inference from noisy, heterogeneous signals. Despite Fido2's fast-growing adoption, no Internet-scale measurement study has characterized the actual cryptographic quality of deployed challenges. This gap matters to the network measurement community, which lacks empirical data on a protocol increasingly central to Web traffic; to website operators, who have no external visibility into the compliance of their own deployment or that of third-party identity providers they depend on; and to end users, whose protection ultimately rests on a link in the authentication chain that has, so far, never been measured in the wild.

This work has been supported by the CyberExcellence project funded by the Walloon Region, under number 2110186.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CoNEXT-SW '26, Utrecht, Netherlands
© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-2976-8/2026/12
https://doi.org/10.1145/3843179.3858058

2 Approach

Atlas addresses this measurement challenge through active, non-intrusive probing of the Fido2 authentication flow. It combines headless browser automation with instrumentation of the WebAuthn API, intercepting `navigator.credentials.get` and `navigator.credentials.create` calls to capture the cryptographic material exchanged during real authentication attempts. To remain within ethical bounds, this capture is performed through passive JavaScript interaction with the page's own API calls, without ever submitting credentials or creating accounts. From each captured exchange, Atlas derives both cryptographic parameters — challenge length, effective entropy, inter-session Hamming distance, and challenge reuse across sessions — and protocol-level configuration parameters, such as the user verification flag and relying party settings.

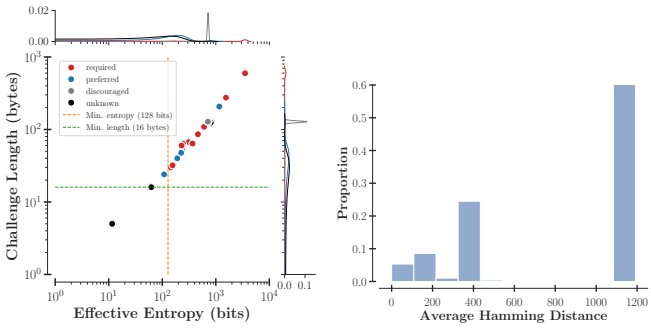
These raw measurements are then condensed into a single, formal conformity score $C(s) = \sum_{k=1}^{4} w_k d_k(s)$, where each dimension $d_k(s) \in [0, 1]$ — challenge quality, user verification policy, relying party configuration, and inter-session robustness — scores a deployment against the corresponding Fido2 specification requirements, and is weighted by its relative security impact ($w_1 = 0.4$, $w_2 = 0.3$, $w_3 = 0.2$, $w_4 = 0.1$). This yields $C(s) \in [0, 1]$, making cross-site and longitudinal comparisons tractable at Internet scale, where inspecting raw parameters site-by-site would not be.

3 Expected Effects, Alternatives & Related Work

Expected Results. The aim is to produce the first large-scale empirical picture of Fido2 deployment quality, characterizing the distribution of conformity scores across sites, sectors, and identity providers, together with a structured taxonomy of vulnerabilities that can be used to discover flawed Fido2 deployments. Such poorly implemented passkey authentication processes may expose websites to replay attacks. Establishing this picture, however, requires positioning Atlas against three lines of work that each capture a different facet of the Fido2 ecosystem, yet stop short of measuring what Atlas targets.

Alternatives. Formal protocol analyses, such as Barbosa et al. [1], establish the theoretical security of the Fido2 protocol itself, but say nothing about whether individual relying parties implement it correctly. At the deployment level, large-scale measurements of passkey adoption [2] reveal *whether* a site supports Fido2, but not *how well*. Closer to the authentication ceremony, the client-side analysis of Kuchal et al. [4] examines authenticator behavior, leaving the server side of the exchange unmeasured.

Contribution. None of these perspectives — protocol, adoption, or client — measures the network-level artifacts actually exchanged during Fido2 flows: the entropy and freshness of the challenge,



(a) Challenges entropy vs. challenges length **(b) Challenges hamming distance**

Figure 1: Preliminary security analysis of Fido2 cryptographic challenges sent by the server ($N = 1,011$).

its reuse across independent sessions, or the negotiated policy parameters. Atlas closes this gap through Internet-scale active measurement of the server side of Fido2 deployments, treating the authentication handshake itself as an observable network protocol whose real-world robustness — not merely its specification — can and should be characterized empirically.

4 Validation Plan

Hypothesis Testing. The central hypothesis is that Fido2 deployments, despite their common reliance on the same underlying protocol, exhibit measurably heterogeneous cryptographic robustness in the wild. To test it, a preliminary characterisation of the overall authentication landscape is first carried out on the Top 1 million Tranco websites [5], providing a broad vantage point on the mechanisms deployed across the Web and identifying a significant population of Fido2-enabled sites. Each identified site is then probed five times across independent authentication sessions — each using a fresh browser profile and separated by a randomized delay — mirroring the repeated-vantage-point methodology common to Internet measurement studies, in order to capture challenge variability across sessions and extract the full set of cryptographic and policy parameters described above.

Reliability. The accuracy of this large-scale characterisation cannot be taken for granted and must itself be measured. A manually constructed groundtruth of 1,000 websites is used to validate the initial categorisation of authentication methods, providing an estimate of classification accuracy and ensuring that the population selected for Fido2 analysis is correctly identified before any cryptographic conclusion is drawn from it.

Plausible Conclusion. All conclusions rely on material directly captured from real authentication flows, rather than on indirect inference or assumptions about protocol behaviour. The full measurement infrastructure and procedures are released as open source, allowing the campaign to be independently reproduced and extended to other vantage points or time periods.

5 Success Metrics & Preliminary Results

Success criteria. The first criterion is the ability to statistically characterise the distribution of Fido2 cryptographic quality in the wild through a conformity score $C(s)$ summarising the measured properties of each deployment. The second is reliable detection of concrete vulnerabilities: exact challenge reuse, identified by comparing the five challenges captured per website, and insufficient entropy, identified against the specification's threshold.

Preliminary results. As a proof of feasibility, 2,575 Fido2-class websites — identified via Atlas's broader classification of the Tranco Top 1M [5] — were probed; 1,011 responded positively to the challenge capture, the rest rejecting the probe due to anti-bot or headless browser detection. Fig. 1a plots the effective entropy (bits), normalized by challenge length to control for the fact that longer challenges naturally admit more bit-level variation, against the challenge length (bytes) for all captured challenges, color-coded by the UserVerification setting. Most challenges cluster above both the minimum entropy (128 bits) and minimum length (16 bytes) thresholds recommended by the specification [7], but a non-negligible fraction — predominantly sites with unknown user verification — fall below one or both, indicating incomplete or careless implementations. An additional analysis concerning the average Hamming distance between the five consecutive cryptographic challenges (Fig. 1b) received for each website highlighted a bimodal structure — most Fido2 implementations generate cryptographically sound challenges, while a minority of sites rely on weak or deterministic generation schemes that undermine the replay-resistance guarantees of the Fido2 specification.

Conclusion. These preliminary results confirm that Atlas reliably captures and analyses real Fido2 cryptographic material at scale, and already surfaces concrete deployment weaknesses. This motivates extending the measurement campaign to the full Top 1 million Tranco websites for a representative, Internet-scale picture of Fido2 cryptographic quality. Scaling this campaign will also require addressing increasingly sophisticated anti-bot and web application firewall (WAF) defenses, which can silently degrade or block captured challenges and must be accounted for to preserve measurement fidelity.

References

[1] M. Barbosa, A. Boldyreva, S. Chen, and B. Warsinchi. 2021. Provable Security Analysis of Fido2. In *Proc. Annual International Cryptology Conference (CRYPTO)*.

[2] P. Bhardwaj and N. Sastry. 2026. State of Passkey Authentication in the Wild: A Census of the Top 100K Sites. In *Proc. Passive and Active Measurement Conference (PAM)*.

[3] J. Hodges, J.C. Jones, M. B. Jones, A. Kumar, E. Lundberg, J. Bradley, C. Brand, A. Langley, G. Mandyam, N. Satragno, N. Steele, J. Tan, S. Weeden, M. West, and J. Yasskin. 2021. *Web Authentication: An API for accessing Public Key Credentials Level 2.* Recommendation 2-20210408. W3C.

[4] D. Kuchal, M. Saa, A. Oest, and F. Li. 2023. Evaluating the Security Posture of Real-World Fido2 Deployments. In *Proc. ACM SIGSAC Conference on Computer and Communications Security (CCS)*.

[5] V. Le Pochat, T. Van Goethem, S. Tajalizadehkhoob, M. Korczynski, and W. Joosen. 2019. Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation. In *Proc. Annual Network and Distributed System Security Symposium (NDSS)*.

[6] Fido Alliance. 2013. Fido2: WebAuthn & CTAP. [Last Accessed: January 12th, 2026]. https://fidoalliance.org/passkeys/

[7] D. Waite, J. Bradley, M. Miller, and T. Cappalli. 2024. *Server Requirements (WebAuthn Level 2 and CTAP 2.1).* Review Draft fido-server-v1.0-rd-20240116. Fido2 Alliance.